Group Expression Intensity Estimation in Videos via Gaussian Processes

Abhinav Dhall[1]
[1]*Australian National University*
abhinav.dhall@anu.edu.au

Roland Goecke[2,1]
[2]*University of Canberra*
roland.goecke@ieee.org

Abstract

Facial expression analysis has been a very active field of research in recent years. This paper proposes a method for finding the apex of an expression, e.g. happiness, in a video containing a group of people based on expression intensity estimation. The proposed method is directly applied to video summarisation based on group happiness and timestamps; further, a novel Gaussian Process Regression based expression intensity estimation method is described. To demonstrate its performance, experiments on smile intensity estimation are performed and compared to other regression based techniques. The smile intensity estimator is extended to group happiness intensity estimation. The proposed intensity estimator can be extended easily for other expressions. The experiments are performed on an 'in the wild' dataset. Quantitative results are presented for comparison of our happiness-intensity detector. A user study was also conducted to verify the results of the proposed method.

1 Introduction

With the large number of movies and millions of online videos (e.g. YouTube) released every year, efficient video summarisation and thumbnail generation for the purpose of video retrieval have become more important than ever. This paper proposes expression apex detection for a group of people using an expression intensity detector. Facial expression analysis is a highly researched topic and much attention has been given to the analysis of an individual's expression in images and videos. However, little attention has been given to expression analysis w.r.t. multiple people in a given frame, nor to the expression intensity. We demonstrate the utility of the proposed approach as a new expression intensity based criterion for video summarisation and evaluate it on the example of expression apex detection for finding the 'happiest group' frame in a video clip, which can be used as a representative frame for the video.

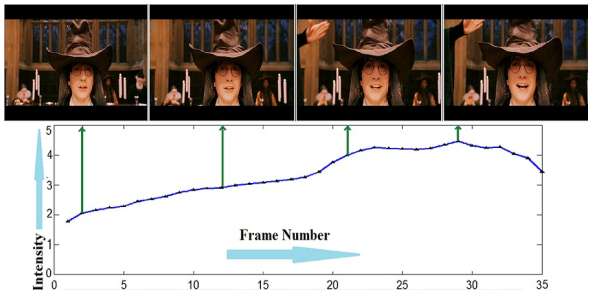

Figure 1. Expression intensity expression estimation on movie data (AFEW [5])

Facial expressions are the visualisation of movement in facial muscles in response to a person's affective state, intentions, or social communications. Automatic facial expression analysis has seen much research over the last two decades and finds many practical applications in human-computer interaction, affective computing, medical problems such as pain and depression analysis, image labelling and retrieval. [12] recently proposed a Support Vector Machine (SVM, [3]) based smile intensity estimator. [15] proposed a new image-based database, GENKI, of smiling and non-smiling images, and evaluated several state-of-the-art methods for happiness detection. However, all of these considered faces independent of each other, not as a group. In contrast, we focus on the group expression.

The main contributions of this paper are: a) expression intensity estimation based on a GPR framework; b) a video summarisation method via face happiness intensity-based clustering and time stamps; c) weighted summation of happiness intensity of multiple subjects in a video frame based on social context.

2 Method

Given a video clip V with N frames r, $V = [r_1, ..., r_N]$ containing multiple actors m, we process

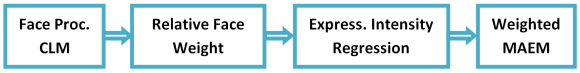

Figure 2. Processing pipeline

each frame as follows: First, face detection [13] is applied. Further, face alignment is performed via non-rigid deformable tracking. The aligned and cropped faces are regressed for predicting their expression intensity. If there are multiple faces in a frame, relative weights are assigned to each face based on their size and distance from the image centroid. Figure 2 shows the processing pipeline for each frame for a video. The proposed framework is evaluated on finding the apex in happy videos. However, this is a generic framework that can be used for other expressions as well.

2.1 Face Processing Pipeline

Saragih *et al.*'s Constrained Local Model (CLM) [11] is used to extract facial landmark points. It is based on fitting a parameterised shape model to the location landmark points of the face. It predicts the locations of the model's landmarks by utilising an ensemble of local feature detectors, which are then combined by enforcing a prior over their joint motion. The distribution of the landmark locations is represented non-parametrically and optimised via subspace constrained meanshifts. The landmark points computed via CLM fitting are used to align the faces.

2.2 Expression Intensity Detection

For a given face f in a frame r_i, we wish to estimate its expression intensity. As discussed earlier, facial expressions are temporal in nature and a basic facial expression comprises of various temporal dynamic stages: *onset*, *apex* and *offset*. As expressions are continuous in nature, we define the following stages for the exemplary smile intensity starting from *neutral* (*onset*) to thrilled (*apex*): *Neutral*, *Small Smile*, *Large Smile*, *Small Laugh*, *Large Laugh* and *Thrilled*.

For the expression intensity estimation problem, given m feature vectors computed from faces, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$, the training set $\mathcal{X}$ is a set of input samples x_i of dimension N and $\mathcal{Y}$ is the corresponding set of vectors y_i of dimension L. The goal here is to learn a mapping function $M : \mathcal{X} \rightarrow \mathcal{Y}$. Gaussian Process Regression (GPR) [9] is used to compute the mapping. GPR has gained increased popularity in statistical machine learning as it offers a principled non-parametric Bayesian framework for inference, model fitting and model selection [1].

In GPR, an output $y_i = f(x_i) + \epsilon_i$ for input x_i and the noise term is assumed to be independent and normally distributed, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2)$. A Gaussian predictive distribution is achieved by adding a Gaussian process prior: $y^*|x^*, \mathcal{D} \sim \mathcal{N}(\mu, \sigma^2)$ with

$$\mu = k^*[K + \sigma_{\mathrm{n}}^2 I]^{-1} y \quad (1)$$

$$\sigma^2 = k(x^*, x^*) + \sigma_{\mathrm{n}}^2 - k^*[K + \sigma_{\mathrm{n}}^2 I]^{-1} k^{*T} \quad (2)$$

for a noisy query point x^*. In these equations, we have $K \in \mathbb{R}^{m \times m}$, $K_{ij} = k(x_i, x_j)$ and $k^* \in \mathbb{R}^{1 \times m}$, $k_i^* = k(x^*, x_i)$. Here, k denotes a covariance function, which encodes the assumptions about the function to be learnt. A squared exponential covariance function of the form

$$k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_p - x_q)^T M_2 (x_p - x_q)\right) + \sigma_n^2 \delta_{pq} \quad (3)$$

is used with $\Theta = \left(\{M_2\}, \sigma_f^2, \sigma_n^2\right)$ that implements automatic relevance determination (ARD) [9], where $\{M_2\}$ denotes the parameters in the symmetric matrix $M_2 = diag(l)^{-2}$. A point prediction y_{guess} can be computed from the Gaussian predictive distribution by minimising the expected loss as

$$y_{\mathrm{opt}}|x^* = \underset{y_{\mathrm{guess}}}{\operatorname{argmin}} \int \mathcal{L}(y^*, y_{\mathrm{guess}}) p(y^*|x^*, \mathcal{D}) dy^* \quad (4)$$

where $\mathcal{L}(.,.)$ is the loss function. For squared loss functions, the point prediction at query point x^* is:

$$y_{\mathrm{opt}}|x^* = \mathbf{E}_{(y^*) \sim p(y^*|x^*, \mathcal{D})}[y^*] = \mu \quad . \quad (5)$$

Figure 1 exemplifies the performance of the proposed smile intensity estimator on a video clip from a facial expressions movie database [5].

Face Descriptors: The input sample vector x_i for each face is computed as follows: The aligned and cropped face (computed in Section 2.1) is used to compute Pyramid of Histogram of Gradients (PHOG) [2] and Local Phase Quantisation (LPQ) [8] features. Parameters for the PHOG descriptor were set as: pyramid level $L = 3$, angle range = $[0 - 360]$ and bin count $= 16$. LPQ is based on computing the short-term Fourier transform on the local image window and has been experimentally shown to better handle blur and illumination than Linear Binary Patterns (LBP) [8]. Then, x_i is a normalised combination of these two histograms. The choice of these two descripors is based on our earlier experiments for emotion recognition [4].

2.3 Handling Multiple Actors

Given a video clip V with N frames r, $V = [r_1, ..., r_N]$ containing multiple actors m, a Multiple

Figure 3. (Left) Input image. (Middle) The smile intensity heat map computed via MAEM. (Right) The smile intensity heat map computed via $\mathrm{MAEM_w}$. It is evident that larger faces get a higher weightage.

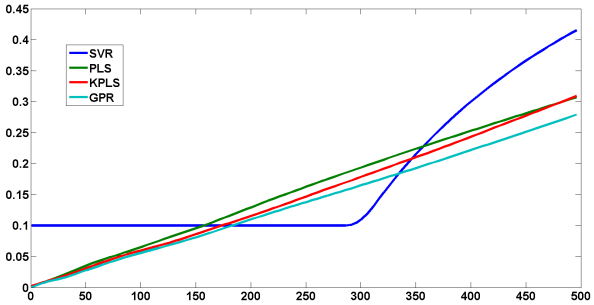

Figure 4. Comparison of the MAEM performance of GPR, KPLS, PLS and SVR.

Actor Expression Model *(MAEM)* can be formulated as an average of the smile intensities for all the faces in a given frame r_i

$$\mathbf{MAEM} = \frac{\sum_i \mathcal{I}_{\mathcal{S}i}}{m} \quad . \tag{6}$$

We wish to include social context information based on the global structure on where actors are located in a given scene. [7] explored social features based on global structures of groups of people in social gathering scenarios. Generally, in group images, people standing close to the camera have relatively larger faces. [6] proposed a framework for detecting the pose of people in a group. They also make a similar assumption to find if a person is standing in the foreground or at the back in a multiple people pose detection scenario.

For the multiple actor setting, we wish to apply weights to the expression intensities of subjects based on the apparent size of their face in the image, as an indicator of the distance to the camera. The size of a face can be estimated by the distance between the location of the eyes given by $s_i = ||\mathbf{l} - \mathbf{r}||$. The relative face size $\theta_\mathbf{i}$ of f_i in frame is then given by

$$\theta_\mathbf{i} = \frac{s_i}{\sum_i s_i/m} \tag{7}$$

where the term $\sum_i s_i/m$ is the mean face size of all the faces in the frame.

Weighted MAEM: The θ_i for each face is further normalised based on the maximum relative distance d_i. Adding this weight to Eq. 6, the weighted MAEM is

$$\mathbf{MAEM}_w = \frac{\sum_i \mathcal{I}_{\mathcal{S}i}\omega_i}{m} \tag{8}$$

where ω_i is the weight defined as $\theta_i/2^{\beta-1}$ and β is a control factor. Figure 3 shows the effect of adding weights to the smile intensities of faces. Left is the input image. The middle shows the smile intensity heat map computed via $\mathbf{MAEM}$ and on the right is the smile intensity heat map computed via $\mathbf{MAEM_w}$. It is evident visually that larger faces get a higher weightage. For a single subject, $\mathbf{MAEM_w}$ is equivalent to $\mathbf{MAEM}$.

3 Experiments and Results

For the experiments, we used the *Acted Facial Expressions in The Wild (AFEW)*[1] [5] database. AFEW is a temporal video database collected from movies based on a semi-automatic subtitle parsing based technique. It is the first facial expression database, which mimics close-to-real-world scenarios and labelled multiple subjects in a clip. The clips in the database have been collected from movies of different genres such as the Harry Potter series and Hugh Grant films, making the database a challenging one.

For the example smile intensity estimation, 3400 faces in Flickr images were manually labelled for their happiness intensity. The GPR based smile intensity estimation method is compared to Kernel Partial Least Squares (KPLS) [10] and Support Vector Regression (SVR) [3]. For KPLS, we set the value of latent factors as 8 and kernel size as 400. For SVR, a Radial Basis Function based kernel and parameters were searched using a five-fold cross validation. Figure 4 describes the comparison on the basis of Mean Average Error (MAE) for smile intensity estimation based on GPR with KPLS and non-linear SVM. The MAE for GPR is 0.7180, for KPLS is 0.9625 , for PLS is 0.9649 and for non-linear SVR is 1.108. We also evaluated the performance of the proposed method on a 'happy' clip from the FEEDTUM [14] database. Figure 6 exemplifies the performance of the smile intensity estimator for the GPR method.

Video Summarisation: $\mathbf{MAEM}_w$ was used for video summarisation. K-means clustering is applied

[1] Available at http://cs.anu.edu.au/few

Figure 5. Top row: Example frames from a video. The two bottom rows show the video summarisation results based on MAEM and $\mathrm{MAEM_w}$ (Sec. 2)

to the frame level happiness intensities. The distances from each point to every centroid are sorted and the frames with the least distance from the centroid are chosen as the summary frames. Further, the selected frames are sorted with respect to their original timestamps. Figure 5 shows the output for a short video clip from AFEW containing multiple subjects. The black boxes show the frames from the video clip, while the middle and the bottom rows show the summarisation frames computed by $\mathbf{MAEM}$ and $\mathbf{MAEM_w}$, respectively. We conducted a user survey, where 10 users were asked to rate on the scale of 0 (not good) - 5 (good), the video summarisation results based on expression for the two methods $\mathbf{MAEM}$ and $\mathbf{MAEM_w}$. A one-way ANOVA analysis ($p < 0.0001$) shows that our hypothesis behind adding weights to intensities holds.

4 Conclusions

The paper proposes a method for estimating the intensity of a facial expression, which we demonstrate on the example of finding the apex of happy expressions in movie clips but which can be similarly applied to other facial expressions. To this end, the smile intensity is estimated based on a Gaussian Process Regression. Further, for handling multiple subjects in a frame, a weighted model based on contextual features is formulated. The contextual feature comprises of the distance to centroid normalised face sizes. We show that the proposed method can be directly applied to expression-based video summarisation. The proposed method can

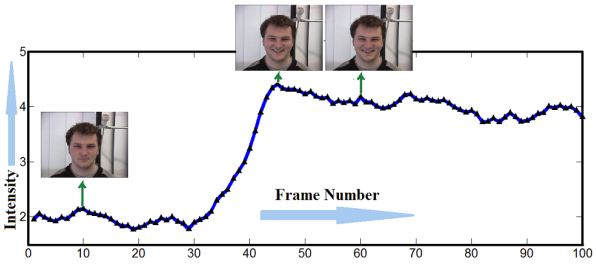

Figure 6. Result for a clip from the FEED-TUM database [14] (Sec. 2)

be easily used with other expressions as well.

References

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer. 2006.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing Shape with a Spatial Pyramid Kernel. In *CIVR*, 2007.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using PHOG and LPQ features. In *IEEE AFGR2011 workshop FERA*, 2011.

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. A semi-automatic method for collecting richly labelled large facial expression databases from movies. *IEEE Multimedia*, 2012.

[6] M. Eichner and V. Ferrari. We Are Family: Joint Pose Estimation of Multiple Persons. In *ECCV*, 2010.

[7] A. Gallagher and T. Chen. Understanding Images of Groups of People. In *CVPR*, 2009.

[8] V. Ojansivu and J. Heikkilä. Blur Insensitive Texture Classification Using Local Phase Quantization. In *ICISP*, pages 236–243, 2008.

[9] C. E. Rasmussen. Gaussian processes for machine learning. MIT Press, 2006.

[10] R. Rosipal. *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*, chapter Nonlinear Partial Least Squares: An Overview. ACCM, IGI Global, 2011.

[11] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 2009.

[12] K. Shimada, T. Matsukawa, Y. Noguchi, and T. Kurita. Appearance-based smile intensity estimation by cascaded support vector machines. In *ACCV Workshops*, 2010.

[13] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[14] F. Wallhoff. Facial expressions and emotion database, 2006. `http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html`.

[15] J. Whitehill, G. Littlewort, I. R. Fasel, M. S. Bartlett, and J. R. Movellan. Toward Practical Smile Detection. *IEEE TPAMI*, 2009.